Base de Datos I
Trabajo Práctico Especial
1er Cuatrimestre 2022

1. *Objetivo*

El objetivo de este Trabajo Práctico Especial es aplicar los conceptos de SQL Avanzado (PSM, Triggers) vistos a lo largo del curso, para implementar funcionalidades y restricciones no disponibles de forma estándar (que no pueden resolverse con Primary Keys, Foreign Keys, etc.).

2. Modalidad

El Trabajo Práctico estará disponible en el Campus a partir del 09/06/2022, indicándose allí mismo la fecha de entrega.
Se incluye junto con el enunciado el archivo: **tourists-rj.csv**.
El TP deberá realizarse en grupos de 3 alumnos, a excepción de un grupo de 4, y entregarse a través de la plataforma Campus ITBA hasta la fecha allí indicada.

3. Descripción del Trabajo

En el sitio de Kaggle se pueden encontrar distintos datasets de acceso público. En esta ocasión vamos a utilizar el dataset de Tourist Arrivals in Rio de Janeiro que consiste en un resumen por año de la cantidad de turistas que hay ingresado a la ciudad de Río de Janeiro entre el 2006 y el 2019. Este análisis ayuda al municipio a identificar de donde provienen sus turistas.

La información es provista en un archivo CSV (Comma Separated Values). El archivo **tourists-rj.csv** contiene información de la cantidad de turistas que ingresaron a la ciudad de Río de Janeiro agrupada por diferentes dimensiones, como ser año, medio de transporte y país (con el detalle de su región y continente). Las columnas del archivo son:

- ***Pais****: nombre del país del cual provienen los turistas ingresados a la ciudad de Río de Janeiro*
- ***Total:*** *cantidad total de turistas ingresados a la ciudad de Río de Janeiro sin importar el medio de transporte utilizado*
- ***Aerea****: cantidad total de turistas ingresados a la ciudad de Río de Janeiro a través de un transporte aereo*
- ***Maritima****: cantidad total de turistas ingresados a la ciudad de Río de Janeiro a través de un transporte marítimo*
- ***Region****: nombre de la región a la cual pertenece el país del cual provienen los turistas ingresados a la ciudad de Río de Janeiro*
- ***Continente****: nombre del continente al cual pertenece el país del cual provienen los turistas ingresados a la ciudad de Río de Janeiro*
- ***Anio****: año en el cual ingresaron los turistas a la ciudad de Río de Janeiro*

Antes de insertar el archivo en una tabla definitiva, se quiere interceptar la inserción del pais, región y continente, y cambiarla por una FK a una dimensión geográfica normalizada en las tablas PAIS, REGION, CONTINENTE. Adicionalmente se quiere interceptar los años y cargar una tabla ANIO.

La finalidad de este Trabajo Práctico Especial consiste en migrar los datos del archivo CSV a una base de datos, producir un reporte y realizar algunas validaciones. Específicamente se debe hacer lo siguiente:

a) Crear las tablas de la dimensión geográfica
b) Crear la tabla para guardar los años
c) Crear la tabla definitiva donde los campos pais, region y continente se reemplazan por un único campo identificatorio de país
d) Importar los datos y cargar la tabla de años y las tablas de la dimensión geográfica
e) Crear un reporte con información consolidada

a) Creación de las tablas de la dimensión geográfica.

Deben crearse las tablas de la dimensión geográfica con las siguientes condiciones:

- CONTINENTE: debe tener un campo identificatorio y campo con el nombre del continente
- REGION: debe tener un campo identificatorio, un campo con el nombre de la región e incluir la FK al CONTINENTE
- PAIS: debe tener un campo identificatorio, un campo con el nombre del país e incluir la FK a la REGION

Se deben crear las claves (pueden ser autogeneradas) y constraints apropiados.

b) Creación de la tabla de años.

Debe crearse la tabla de años la cual debe incluir un campo que indique si el año es bisiesto o no.

c) Creación de la tabla definitiva.

Debe crearse un tabla definitiva que será la receptora de los datos del archivo **tourists-rj.csv**. Los campos y restricciones de la tabla deben crearse en base al análisis de los datos. Recordar que los archivos csv son archivos de texto que pueden abrirse fácilmente con cualquier editor. Se recomienda que los nombres de los campos coincidan con los de las columnas del archivo csv.

Para el caso particular de los campos pais, region y continente, se deberá cambiar su contenido para que el mismo haga referencia a la key de la tabla PAIS antes de insertarlo en la tabla definitiva.

En base a los datos, se debe crear la clave y constraints apropiados.

d) Importación de los datos.

Utilizando el comando COPY de PostgreSQL, se deben importar TODOS los datos del archivo **csv** en la tabla creada en **c)**. El archivo **csv** provisto por la cátedra NO puede ser modificado.

Creación de un trigger para:

1) Determinar la FK de la dimensión geográfica

Para insertar los datos en la tabla definitiva es necesario interceptar la inserción del pais, region y contintente, y agregar la FK a la dimensión geográfica de la tabla PAIS.

2) Cargar los valores de la estructura de la dimensión geográfica

Además de insertar los datos del archivo, se deben poblar las distintas tablas que conforman la dimensión geográfica, siempre y cuando los valores correspondientes no existan en dichas tablas.

Por ejemplo, si partimos con las tablas de la dimensión geográficas vacías y si al principio en el archivo CSV viene el pais ”Argentina”, con la region ”América do Sul” y con el continente “América“, se debe insertar una tupla en cada una de las tablas de geografía, quedando las tablas con la siguiente información:

• CONTINENTE: ”América”

• REGION: ”América do Sul”, FK al continente ”América”

• PAIS: ”Argentina”, FK a la región ”América do Sul”

Sin embargo, si luego viene el pais ”Chile”, con la region ”América do Sul” y con el continente “América“, se reaprovecha la región y el continente ya cargados, quedando las 2 tablas REGION y CONTINENTE igual que antes. Entonces solamente se debe insertar una tupla en PAIS, agregando en la tabla la siguiente información:

• PAIS: ”Chile”, FK a la región ”América do Sul”

Si luego viene nuevamente información de otro año para el pais ”Argentina”, con la region ”América do Sul” y con el continente “América“, no se tiene que insertar ninguna tupla.

Las tuplas en las tablas tienen que tener todos los datos bien completos.

3) Cargar los valores de los años

Además de insertar los datos del archivo, se debe poblar la tabla para guardar los años, siempre y cuando los mismos no existan en dicha tabla.

Por ejemplo, si partimos con la tabla vacía y si al principio en el archivo CSV viene el año 2006, se debe insertar una tupla quedando la tabla con la siguiente información:

• ANIO: 2006, bisiesto = false

Si luego viene el año ”2008” se debe insertar una tupla en ANIO, agregando en la tabla la siguiente información:

• ANIO: 2008, bisiesto = true

Sin embargo, si luego viene nuevamente información de ”2008”, no se tiene que insertar ninguna tupla. Las tuplas en la tabla tiene que tener todos los datos bien completos.

e) Reporte de información consolidada.

El intendente de Río de Janeiro realiza un análisis de los turistas con información consolidada, agrupada por Año y por distintas categorías como ser Continente y Medio de transporte.

Se pide crear la función **AnalisisConsolidado(n)** que recibe como parámetro la cantidad de años a mostrar tomando como base el primer año cargado en la tabla definitiva, la cual genere un reporte mostrando para cada Año y categoría, la cantidad total de turistas y el promedio.

El reporte tendrá las siguientes características:

I. Título del reporte:

```
"CONSOLIDATED TOURIST REPORT”
```

II. Encabezado de columnas:

```
“Year   Category    Total       Average”
```

III. Por cada año tiene que aparecer un renglón en el reporte, con el Año ordenados de menor a mayor. La primer categoría de agrupación (Continente) con sus valores ordenados alfabéticamente y sus métricas (Total y Average), deben estar en el **mismo** renglón que el año y el resto de las categorías (Medio de Transporte), encolumnados a continuación en los renglones subsiguientes

IV. Al final de todo, tiene que aparecer el total de las métricas Total y Average correspondientes para ese año

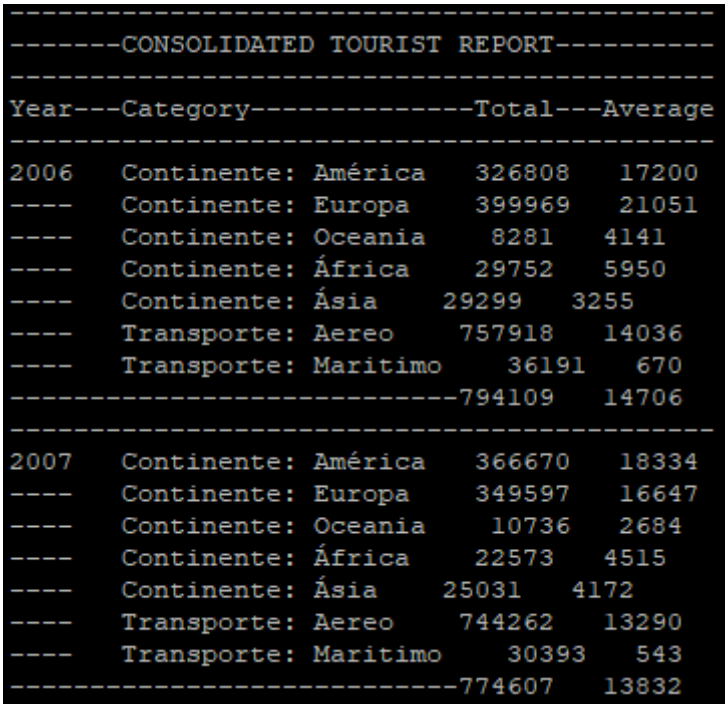
En caso de que no existieran datos para los parámetros ingresados, no se debe mostrar nada (ni siquiera el encabezado).
La función debe manejar los posibles errores.

Por ejemplo,

- si invocamos **AnalisisConsolidado(1)** se debe obtener información del año 2006:

```
-----------------------------------------
-------CONSOLIDATED TOURIST REPORT----------
-----------------------------------------
Year---Category-------------Total---Average
-----------------------------------------
2006   Continente: América   326808   17200
----   Continente: Europa    399969   21051
----   Continente: Oceania     8281    4141
----   Continente: África    29752    5950
----   Continente: Ásia    29299   3255
----   Transporte: Aereo    757918   14036
----   Transporte: Maritimo    36191    670
---------------------------794109   14706
```

- Si invocamos **AnalisisConsolidado(2)** se debe obtener información del año 2006 y 2007



```
-------------------------------------------
-------CONSOLIDATED TOURIST REPORT----------
-------------------------------------------
Year---Category-------------Total---Average
-------------------------------------------
2006   Continente: América   326808   17200
----   Continente: Europa    399969   21051
----   Continente: Oceania    8281   4141
----   Continente: África    29752   5950
----   Continente: Ásia    29299   3255
----   Transporte: Aereo    757918   14036
----   Transporte: Maritimo    36191   670
---------------------------794109   14706
-------------------------------------------
2007   Continente: América   366670   18334
----   Continente: Europa    349597   16647
----   Continente: Oceania    10736   2684
----   Continente: África    22573   4515
----   Continente: Ásia    25031   4172
----   Transporte: Aereo    744262   13290
----   Transporte: Maritimo    30393   543
---------------------------774607   13832
```

- Si invocamos **AnalisisConsolidado(15)** se debe obtener obtener lo mismo que invocando **AnalisisConsolidado(14)**, es decir información del año 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018 y 2019

- Si invocamos **AnalisisConsolidado(0)** no se obtiene nada

4. Entregables

Los alumnos deberán entregar los siguientes documentos:

- El script sql **funciones.sql** con el código necesario para crear las tablas, las funciones y los triggers

- Un informe que debe contener:
    - El rol de cada uno de los participantes del grupo. Si bien en el TP deben estar involucrados todos los integrantes, se debe asignar un rol de supervisión de cada una de las tareas. Mínimamente los roles son: encargado del informe, encargado de las funciones, encargado del trigger, encargado del funcionamiento global del proyecto y encargado de investigación. Pueden asignarse más roles en caso de requerirse
    - Todo lo investigado para realizar el TP
    - Las dificultades encontradas y cómo se resolvieron
    - También se debe detallar aquí el proceso de importación de los datos realizado
    - El informe debe tener como máximo 3 páginas

5. Evaluación

La evaluación del trabajo se llevará a cabo utilizando los parámetros establecidos en la rúbrica asociada a la actividad en el Campus.
Se tendrá en cuenta que las consultas, más allá del funcionamiento (lo cual es fundamental), sean genéricas.
Los docentes ejecutarán el proceso usando los conjuntos de datos entregados pero podrán también hacer pruebas con otros conjuntos de datos de similares características para evaluar el funcionamiento en distintos escenarios.
El informe deberá estar completo y sin faltas de ortografía.
En caso de que el trabajo no cumpliera los requisitos básicos para ser aprobado, los alumnos serán citados en la fecha de recuperatorio para defenderlo y corregir los errores detectados.